WEST

Generate Collection

L4: Entry 1 of 6 File: USPT Jan 26, 1999

DOCUMENT-IDENTIFIER: US 5863722 A
TITLE: Method of sorting polynucleotides

BSPR:
In one aspect of my invention, tag complements attached to a solid phase support are used to sort polynucleotides from a mixture of polynucleotides each containing a tag. In this embodiment, tag complements are synthesized on the surface of a solid phase support, such as a microscopic bead or a specific location in an array of synthesis locations on a single support, such that populations of identical sequences are produced in specific regions. That is, the surface of each support, in the case of a bead, or of each region, in the case of an array, is derivatized by only one type of tag complement which has a particular sequence. The population of such beads or regions contains a repertoire of tag complements with distinct sequences, the size of the repertoire depending on the number of subunits per oligonucleotide tag and the length of the subunits employed, where oligomeric subunits are used. Similarly, the polynucleotides to be sorted each comprises an oligonucleotide tag in the repertoire, such that identical polynucleotides have the same tag and different polynucleotides have different tags. Thus, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, subpopulations of identical polynucleotides are sorted onto particular beads or regions. The subpopulations of polynucleotides can then be manipulated on the solid phase support by micro-biochemical techniques.

BSPR:
An important aspect of my invention is the use of the oligonucleotide tags to sort polynucleotides for parallel sequence determination. Preferably, this aspect of the invention comprises the following steps: (a) generating from a target polynucleotide a plurality of fragments that covers the target polynucleotide; (b) attaching an oligonucleotide tag from a repertoire of tags to each fragment of the plurality (i) such that substantially all the same fragments have the same oligonucleotide tag attached and (ii) such that each oligonucleotide tag from the repertoire comprises a plurality of subunits and each subunit of the plurality consists of a complementary nucleotide of an antisense monomer or an oligonucleotide having a length from three to six nucleotides, the oligonucleotides being selected from a minimally cross-hybridizing set; (c) sorting the fragments by specifically hybridizing the oligonucleotide tags with their respective tag complements; (d) determining the nucleotide sequence of a portion of each of the fragments of the plurality; and (e) determining the nucleotide sequence of the target polynucleotide by collating the sequences of the fragments.

BSPR:
Another important feature of my invention is a method of identifying, or fingerprinting, a population of mRNA molecules. Preferably, such a method comprises the following steps: (a) forming a population of cDNA molecules from the population of mRNA molecules, the cDNA molecules being complementary to the mRNA molecules and each cDNA molecule having an oligonucleotide tag attached, (i) such that substantially all of the same cDNA molecules have the same oligonucleotide tag attached and (ii) such that each oligonucleotide tag from the repertoire comprises a plurality of subunits and each subunit of the plurality consists of a complementary nucleotide of an antisense monomer or an

oligonucleotide having a length from three to six nucleotides, the oligonucleotides being selected from a minimally cross-hybridizing set; (b) sorting the cDNA molecules by specifically hybridizing the oligonucleotide tags with their respective tag complements; (c) determining the nucleotide sequence of a portion of each of the sorted cDNA molecules; and (d) identifying the population of mRNA molecules by the frequency distribution of the portions of sequences of the cDNA molecules.

DEPR:
The invention provides a method of labeling and sorting molecules, particularly polynucleotides, by the use of oligonucleotide tags. In one aspect, the oligonucleotide tags of the invention comprise a plurality of "words" or subunits selected from minimally cross-hybridizing sets of subunits. Subunits of such sets cannot form a duplex or triplex with the complement of another subunit of the same set with less than two mismatched nucleotides. Thus, the sequences of any two oligonucleotide tags of a repertoire that form duplexes will never be "closer" than differing by two nucleotides. In particular embodiments, sequences of any two oligonucleotide tags of a repertoire can be even "further" apart, e.g. by designing a minimally cross-hybridizing set such that subunits cannot form a duplex with the complement of another subunit of the same set with less than three mismatched nucleotides, and so on. Usually, oligonucleotide tags of the invention and their complements are oligomers of the natural nucleotides so that they may be conveniently processed by enzymes, such as ligases, polymerase, terminal transferases, and the like.

DEPR:
Solid phase supports for use with the invention may have a wide variety of forms, including microparticles, beads, and membranes, slides, plates, micromachined chips, and the like. Likewise, solid phase supports of the invention may comprise a wide variety of compositions, including glass, plastic, silicon, alkanethiolate-derivatized gold, cellulose, low cross-linked and high cross-linked polystyrene, silica gel, polyamide, and the like. Preferably, either a population of discrete particles are employed such that each has a uniform coating, or population, of complementary sequences of the same tag (and no other), or a single or a few supports are employed with spacially discrete regions each containing a uniform coating, or population, of complementary sequences to the same tag (and no other). In the latter embodiment, the area of the regions may vary according to particular applications; usually, the regions range in area from several .mu.m.sup.2, e.g. 3-5, to several hundred .mu.m.sup.2, e.g. 100-500. Preferably, such regions are spacially discrete so that signals generated by events, e.g. fluorescent emissions, at adjacent regions can be resolved by the detection system being employed. In some applications, it may be desirable to have regions with uniform coatings of more than one tag complement, e.g. for simultaneous sequence analysis, or for bringing separately tagged molecules into close proximity.

DEPR:
An important aspect of the invention is the sorting of populations of identical polynucleotides, e.g. from a cDNA library, and their attachment to microparticles or separate regions of a solid phase support such that each microparticle or region has only a single kind of polynucleotide. This latter condition can be essentially met by ligating a repertoire of tags to a population of polynucleotides followed by cloning and sampling of the ligated sequences. A repertoire of oligonucleotide tags can be ligated to a population of polynucleotides in a number of ways, such as through direct enzymatic ligation, amplification, e.g. via PCR, using primers containing the tag sequences, and the like. The initial ligating step produces a very large population of tag-polynucleotide conjugates such that a single tag is generally attached to many different polynucleotides. However, by taking a sufficiently small sample of the conjugates, the probability of obtaining "doubles," i.e. the same tag on two different polynucleotides, can be made negligible. (Note that it is also possible to obtain different tags with the same polynucleotide in a sample. This case is simply leads to a polynucleotide

being processed, e.g. sequenced, twice). As explain more filly below, the probability of obtaining a double in a sample can be estimated by a Poisson distribution since the number of conjugates in a sample will be large, e.g. on the order of thousands or more, and the probability of selecting a particular tag will be small because the tag repertoire is large, e.g. on the order of tens of thousand or more. Generally, the larger the sample the greater the probability of obtaining a double. Thus, a design trade-off exists between selecting a large sample of tag-polynucleotide conjugates--which, for example, ensures adequate coverage of a target polynucleotide in a shotgun sequencing operation, and selecting a small sample which ensures that a minimal number of doubles will be present. In most embodiments, the presence of doubles merely adds an additional source of noise or, in the case of sequencing, a minor complication in scanning and signal processing, as microparticles giving multiple fluorescent signals can simply be ignored. As used herein, the term "substantially all" in reference to attaching tags to molecules, especially polynucleotides, is meant to reflect the statistical nature of the sampling procedure employed to obtain a population of tag-molecule conjugates essentially free of doubles. The meaning of substantially all in terms of actual percentages of tag-molecule conjugates depends on how the tags are being employed. Preferably, for nucleic acid sequencing, substantially all means that at least eighty percent of the tags have unique polynucleotides attached. More preferably, it means that at least ninety percent of the tags have unique polynucleotides attached. Still more preferably, it means that at least ninety-five percent of the tags have unique polynucleotides attached. And, most preferably, it means that at least ninety-nine percent of the tags have unique polynucleotides attached.

DEPR:
Preferably, when the population of polynucleotides is messenger RNA (mRNA), oligonucleotides tags are attached by reverse transcribing the mRNA with a set of primers containing complements of tag sequences. An exemplary set of such primers could have the following sequence (SEQ ID NO: 1):

DEPR:
The above method may be used to fingerprint mRNA populations when coupled with the parallel sequencing methodology described below. Partial sequence information is obtained simultaneously from a large sample, e.g. ten to a hundred thousand, of cDNAs attached to separate microparticles as described in the above method. The frequency distribution of partial sequences can identify mRNA populations from different cell or tissue types, as well as from diseased tissues, such as cancers. Such mRNA fingerprints are useful in monitoring and diagnosing disease states.

DEPC:
Parallel Sequencing

DEPC:
Parallel Sequencing of SV40 Fragments

CLPR:
1. A method of characterizing a population of mRNA molecules, the method comprising the steps of:

CLPR:
2. The method of claim 1 wherein said one or more solid phase supports are a plurality of microparticles.

CLPR:
5. The method of claim 1 wherein said one or more solid phase supports are a plurality of microparticles.

CLPR:
6. The method of claim 5 wherein after said step of sorting, said plurality of microparticles is fixed to a planar substrate.

CLPR:
7. The method of claim 6 wherein said plurality of microparticles are disposed randomly on the surface of said planar substrate at a density of between 1000 microparticles to 100 thousand microparticles per square centimeter.

CLPR:
9. The method of claim 1 wherein said population of cDNA molecules contains from 500 to 1000 cDNA molecules.

CLPR:
10. A method of characterizing a population of mRNA molecules, the method comprising the steps of:

CLPR:
12. The method of claim 11 wherein said step of determining said nucleotide sequence of said cDNA molecules is carried out simultaneously for said population of cDNA molecules by a single base sequencing method.

CLPV:
forming a population of cDNA molecules from the population of mRNA molecules, the cDNA molecules being complementary to the mRNA molecules and each cDNA molecule having an oligonucleotide tag attached, (i) such that substantially all different cDNA molecules have different oligonucleotide tags attached and (ii) such that each oligonucleotide tag comprises a plurality of subunits and each subunit of the plurality consists of an oligonucleotide having a length from three to six nucleotides or from three to six basepairs, the subunits being selected from a minimally cross-hybridizing set wherein a subunit of the set and a complement of any other subunit of the set would have at least two mismatches,

CLPV:
characterizing the population of mRNA molecules by the frequency distribution of the portions of sequences of the cDNA molecules.

CLPV:
forming a population of cDNA molecules from the population of mRNA molecules, the cDNA molecules being complementary to the mRNA molecules;

CLPV:
attaching an oligonucleotide tag from a repertoire of tags to each cDNA molecule of the population such that substantially all different cDNA molecules have different oligonucleotide tags attached;

CLPV:
determining the nucleotide sequence of a portion of each of the cDNA molecules of the population; and

CLPV:
characterizing the population of mRNA molecules by the frequency distribution of the portions of sequences of the cDNA molecules.